Austin Schwinn
Data Mining
Final Project
March 24, 2017

Github Repository

Data Mining Greenhouse Gases and Global Average Temperature

Introduction:

When I originally started this project a few weeks ago, I thought I would look at trying to identify different global temperature trends by regions. However, about half way through my project I stumbled upon a data science competition through the website KDNuggets.com in conjunction with Data.World and Data4Democracy (a group I have started volunteering for). The competition was entitled Data Science vs Fake News and the goal was to use data visualization to disprove a fake news statement. 2 days before the competition was due, the Administrator for the US Environmental Protection Agency, Scott Pruitt, claimed in a news interview that CO2 is not a major driver in climate change. This gave me the opportunity to use what I had been working on for this project in the competition as well. Attached to my project is my article entry to the competition. Just earlier this week, I learned that my submission earned 2nd place. On top of the data visualization that I used as the basis for my article, I decided to develop my analysis further for this semester project. For my Data Mining extension, I went beyond visualizing current trends to predicting future trends as well with linear regressions. This is all outlined in this report and in my R code.

1. Problem Understanding

This analysis is based around climate change and CO2 as its root cause to disprove Mr. Scott Pruitt's statements. To do this, we will look at how current trends are broken from historical trends. We will also how CO2 and global temperature are correlated and the physical properties of why. Additionally, I want to project what will happen if we continue to follow current trends. Before doing this project, I spent a large amount of time just studying the science behind climate change and how we know what we know. This background knowledge gave me the

understanding of what exactly Scott Pruitt was incorrect.

2. Data Understanding

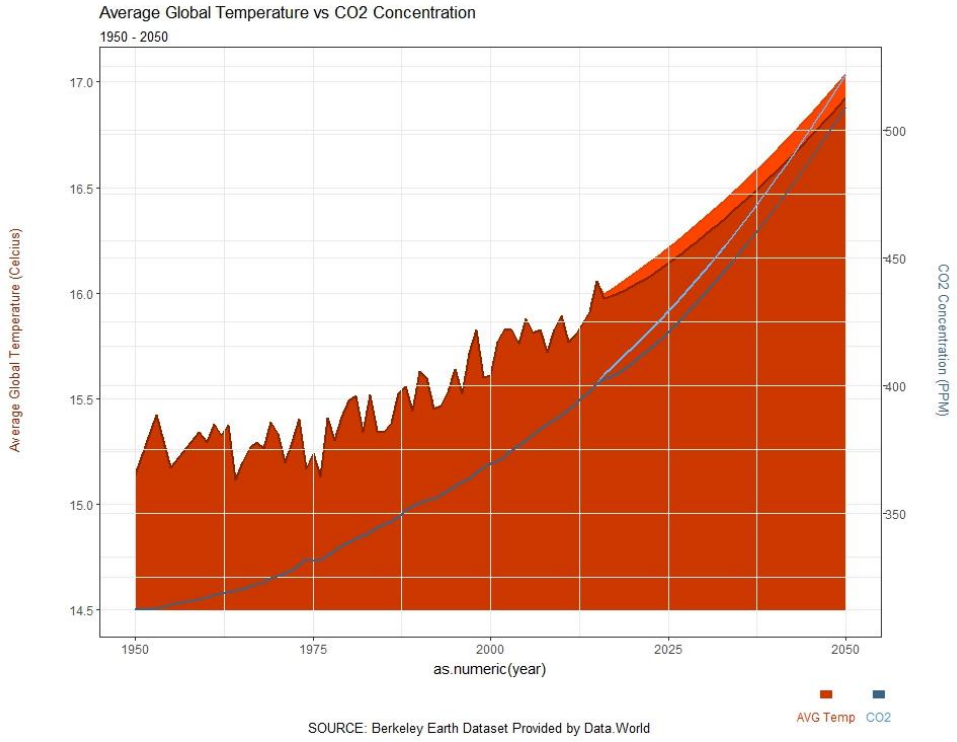
As I spent a large amount of time researching the topic, I also spent a lot of time trying to find the correct datasets. On top of the Berkley Earth Global Average Temperature Data, which was recommended to me, I wanted to find data which was the cause to the temperature change. Two US organizations where particularly helpful, the National Oceanic And Atmospheric Administration (NOAA) and the EPA (which Mr. Pruitt is ironically in charge of). Our pre-historic gas records are important because they shed light on how the cycle has changed in more recent times. The radiative forcing data is useful because it tells us exactly how much each gas affects the temperature gain. And then our temperature information tells us the impact of the other data we fed.

3. Data Preparation

Much of the data preparation step was spent cleaning the data. I developed a function to clean the gas concentration data from the EPA as there were missing values, description headers, and stitched together 10 different data sources into 1. The radiative forcing data was very easy to prepare as it was well organized to being with. The Berkley Earth data was the easiest to prepare as it was from a Kaggle competition. I also had to combine multiple sources like the temperature and co2 data. I also subset my data into training and testing sets for regression fitting and testing later.

4. Modeling

Most of my modeling took place as data visualizations which I used as support for my article. I also used linear regression to find the regression line for CO2 over time. I then ran a second regression to see how CO2 concentration affected the rise in temperature over time. I had a second set of regressions based on CO2 if the rate of change was reduced by a user specified percent. I did this so that someone can play with how much impact we can have on the outcome of the climate based on our change.



5. Evaluation

As I mentioned earlier, I broke my regression data into testing and training sets to see how well my fit my regression lines were. This became an important step as I had to alter my time series linear regression as it follows exponential growth. I eventually solved the issue and the regressions are performing as they should.

6. Deployment

My main deployment was my article that I wrote. I also developed a chart to see how changing the rate of change of CO2 growth would change the rise in temperature over time. I would eventually like to take this idea and implement it in a shiny app where the user can alter the rate of change and see the updated graph in real time.